Matriz de covarianza

Vimos el concepto de co-varianza y la cantidad relacionada que es el coeficiente de correlación, el cual es un numero que mide el grado de relación entre las variaciones de una variable con otra.

¿Qué pasa en los casos donde tengo una aplicación de muchas variables? Como en los data sets. Lo que deberíamos hacer es poder calcular las posibles parejas de todas las co-varianzas que hay en mi data set, el resultado de calcular todos esos valores y cuando lo arreglamos en una matriz es: **LA MATRIZ DE CO-VARIANZA**.

Vamos a ver como se define y construye esta matriz.

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Con esta forma yo puedo saber que puede suceder con un dataset con muchas variables, hay varias columnas, supongamos todas numéricas.

La voy a denotar como:

$$\Sigma = \text{Matriz de co-varianza}$$

Cuando medimos la co-varianza de una variable consigo misma; hablamos de la desviación estándar

$$\Sigma = \begin{array}{c|cccc|} & x & y & z & \dots \\ x & \sigma(x)^2 & c(x,y) & c(x,z) & c(x,\dots) \\ y & c(y,x) & \sigma(y)^2 & c(y,z) & c(y,\dots) \\ z & c(z,x) & c(z,y) & \sigma(z)^2 & c(z,\dots) \\ \vdots & c(\dots) & c(\dots) & c(\dots) & \sigma(\dots)^2 \end{array}$$

In [ ]:

```
#Preparando datos
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler

#Importando un data set
iris=sns.load_dataset('iris')
```

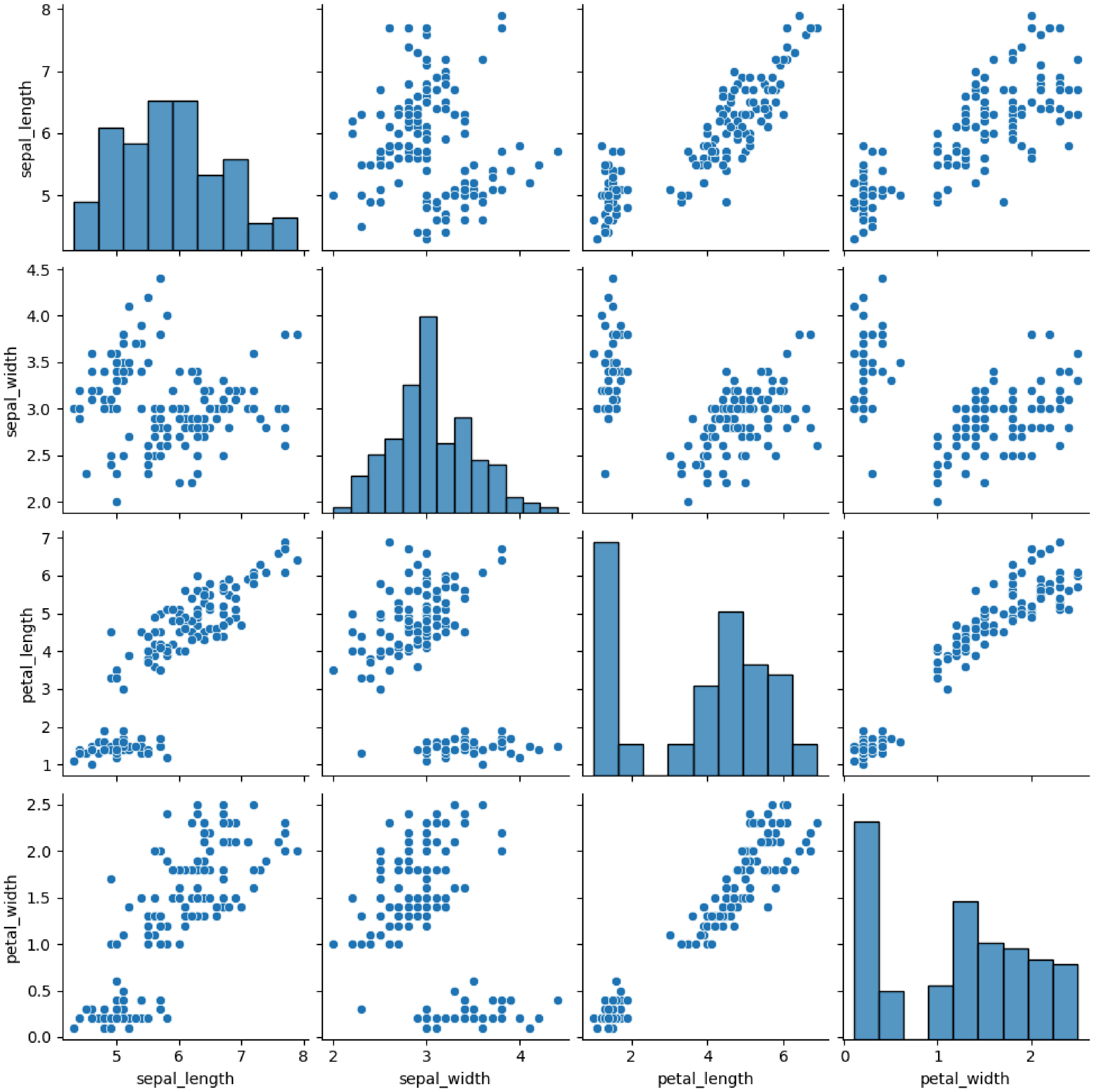
Vamos a aplicar la matriz de correlación

In [ ]:

```
#Obteniendo la matriz visualmente
sns.pairplot(iris)
```

Out[ ]:

```
<seaborn.axisgrid.PairGrid at 0x7fec7cf6c200>
```

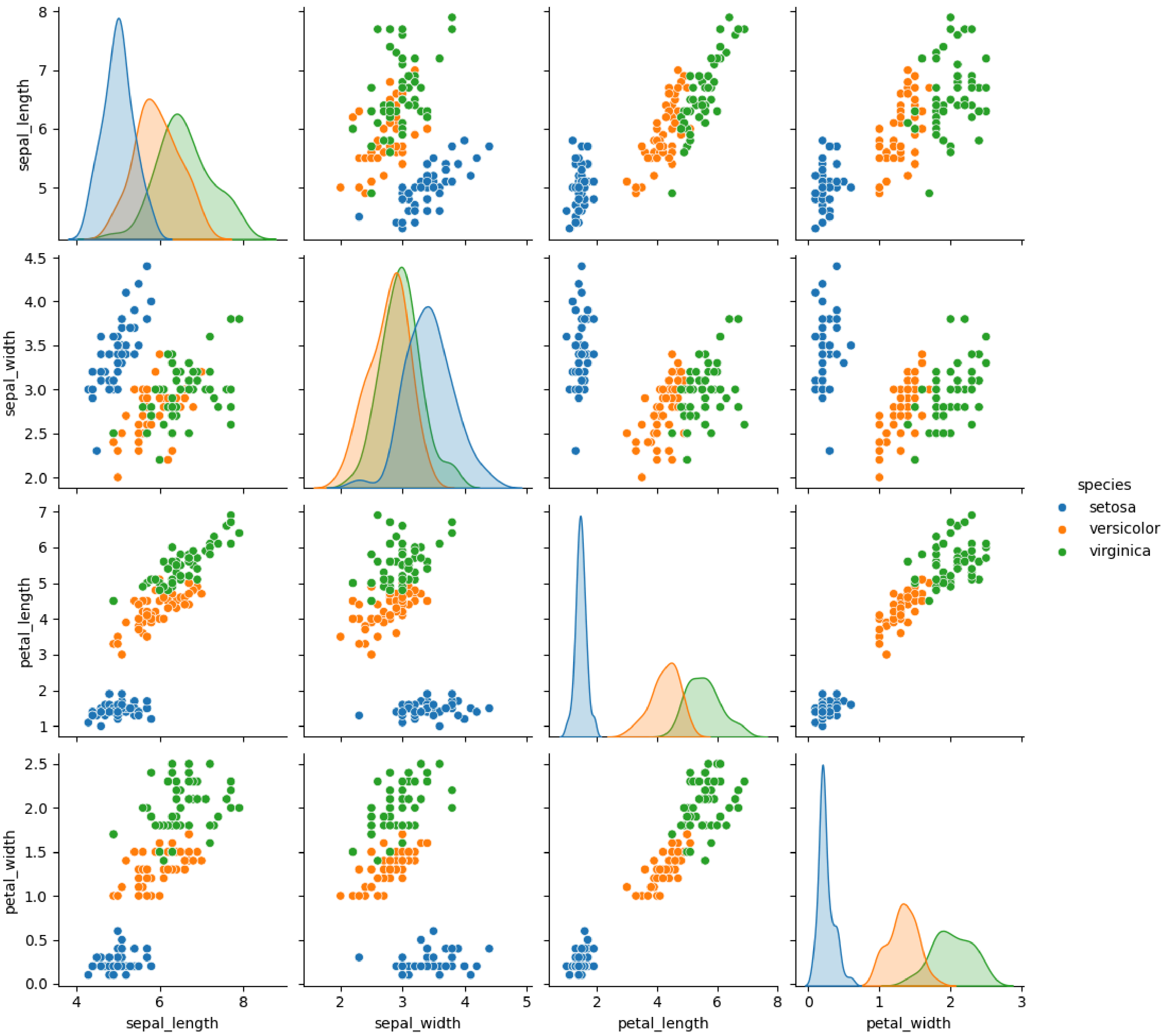


Vemos en la diagonal central, como se relaciona consigo misma, así que Seaborn nos muestra un histograma, ya que no tiene sentido graficar una variable consigo misma porque daría una linea recta.

Hay que categorizarlo con species.

```
In [ ]: #Obteniendo la matriz visualmente
        sns.pairplot(iris,hue='species')
```

```
Out[ ]: <seaborn.axisgrid.PairGrid at 0x7fec0183c8c0>
```



Podemos hacer una correlación si cumple con ciertos grupo o no, gracias al **hue**

In [ ]:
```
iris.columns
```

Out[ ]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

In [ ]:
```
# Creando una normalización
scaler = StandardScaler()

# sin columna de species
data_scaled=scaler.fit_transform(
iris[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']])
data_scaled
```

```
Out[ ]:  array([[-9.00681170e-01,  1.01900435e+00, -1.34022653e+00,
         -1.31544430e+00],
        [-1.14301691e+00, -1.31979479e-01, -1.34022653e+00,
         -1.31544430e+00],
        [-1.38535265e+00,  3.28414053e-01, -1.39706395e+00,
         -1.31544430e+00],
        [-1.50652052e+00,  9.82172869e-02, -1.28338910e+00,
         -1.31544430e+00],
        [-1.02184904e+00,  1.24920112e+00, -1.34022653e+00,
         -1.31544430e+00],
        [-5.37177559e-01,  1.93979142e+00, -1.16971425e+00,
         -1.05217993e+00],
        [-1.50652052e+00,  7.88807586e-01, -1.34022653e+00,
         -1.18381211e+00],
        [-1.02184904e+00,  7.88807586e-01, -1.28338910e+00,
         -1.31544430e+00],
        [-1.74885626e+00, -3.62176246e-01, -1.34022653e+00,
         -1.31544430e+00],
        [-1.14301691e+00,  9.82172869e-02, -1.28338910e+00,
         -1.44707648e+00],
        [-5.37177559e-01,  1.47939788e+00, -1.28338910e+00,
         -1.31544430e+00],
        [-1.26418478e+00,  7.88807586e-01, -1.22655167e+00,
         -1.31544430e+00],
        [-1.26418478e+00, -1.31979479e-01, -1.34022653e+00,
         -1.44707648e+00],
        [-1.87002413e+00, -1.31979479e-01, -1.51073881e+00,
         -1.44707648e+00],
        [-5.25060772e-02,  2.16998818e+00, -1.45390138e+00,
         -1.31544430e+00],
        [-1.73673948e-01,  3.09077525e+00, -1.28338910e+00,
         -1.05217993e+00],
        [-5.37177559e-01,  1.93979142e+00, -1.39706395e+00,
         -1.05217993e+00],
        [-9.00681170e-01,  1.01900435e+00, -1.34022653e+00,
         -1.18381211e+00],
        [-1.73673948e-01,  1.70959465e+00, -1.16971425e+00,
         -1.18381211e+00],
        [-9.00681170e-01,  1.70959465e+00, -1.28338910e+00,
         -1.18381211e+00],
        [-5.37177559e-01,  7.88807586e-01, -1.16971425e+00,
         -1.31544430e+00],
        [-9.00681170e-01,  1.47939788e+00, -1.28338910e+00,
         -1.05217993e+00],
        [-1.50652052e+00,  1.24920112e+00, -1.56757623e+00,
         -1.31544430e+00],
        [-9.00681170e-01,  5.58610819e-01, -1.16971425e+00,
         -9.20547742e-01],
        [-1.26418478e+00,  7.88807586e-01, -1.05603939e+00,
         -1.31544430e+00],
        [-1.02184904e+00, -1.31979479e-01, -1.22655167e+00,
         -1.31544430e+00],
        [-1.02184904e+00,  7.88807586e-01, -1.22655167e+00,
         -1.05217993e+00],
        [-7.79513300e-01,  1.01900435e+00, -1.28338910e+00,
         -1.31544430e+00],
        [-7.79513300e-01,  7.88807586e-01, -1.34022653e+00,
         -1.31544430e+00],
        [-1.38535265e+00,  3.28414053e-01, -1.22655167e+00,
         -1.31544430e+00],
        [-1.26418478e+00,  9.82172869e-02, -1.22655167e+00,
         -1.31544430e+00],
        [-5.37177559e-01,  7.88807586e-01, -1.28338910e+00,
         -1.05217993e+00],
        [-7.79513300e-01,  2.40018495e+00, -1.28338910e+00,
         -1.44707648e+00],
        [-4.16009689e-01,  2.63038172e+00, -1.34022653e+00,
         -1.31544430e+00],
        [-1.14301691e+00,  9.82172869e-02, -1.28338910e+00,
         -1.31544430e+00],
        [-1.02184904e+00,  3.28414053e-01, -1.45390138e+00,
         -1.31544430e+00],
        [-4.16009689e-01,  1.01900435e+00, -1.39706395e+00,
         -1.31544430e+00],
        [-1.14301691e+00,  1.24920112e+00, -1.34022653e+00,
         -1.44707648e+00],
        [-1.74885626e+00, -1.31979479e-01, -1.39706395e+00,
         -1.31544430e+00],
        [-9.00681170e-01,  7.88807586e-01, -1.28338910e+00,
         -1.31544430e+00],
        [-1.02184904e+00,  1.01900435e+00, -1.39706395e+00,
         -1.18381211e+00],
        [-1.62768839e+00, -1.74335684e+00, -1.39706395e+00,
         -1.18381211e+00],
        [-1.74885626e+00,  3.28414053e-01, -1.39706395e+00,
         -1.31544430e+00],
        [-1.02184904e+00,  1.01900435e+00, -1.22655167e+00,
         -7.88915558e-01],
        [-9.00681170e-01,  1.70959465e+00, -1.05603939e+00,
         -1.05217993e+00],
[-1.26418478e+00, -1.31979479e-01, -1.34022653e+00,
 -1.18381211e+00],
[-9.00681170e-01,  1.70959465e+00, -1.22655167e+00,
 -1.31544430e+00],
[-1.50652052e+00,  3.28414053e-01, -1.34022653e+00,
 -1.31544430e+00],
[-6.58345429e-01,  1.47939788e+00, -1.28338910e+00,
 -1.31544430e+00],
[-1.02184904e+00,  5.58610819e-01, -1.34022653e+00,
 -1.31544430e+00],
[ 1.40150837e+00,  3.28414053e-01,  5.35408562e-01,
  2.64141916e-01],
[ 6.74501145e-01,  3.28414053e-01,  4.21733708e-01,
  3.95774101e-01],
[ 1.28034050e+00,  9.82172869e-02,  6.49083415e-01,
  3.95774101e-01],
[-4.16009689e-01, -1.74335684e+00,  1.37546573e-01,
  1.32509732e-01],
[ 7.95669016e-01, -5.92373012e-01,  4.78571135e-01,
  3.95774101e-01],
[-1.73673948e-01, -5.92373012e-01,  4.21733708e-01,
  1.32509732e-01],
[ 5.53333275e-01,  5.58610819e-01,  5.35408562e-01,
  5.27406285e-01],
[-1.14301691e+00, -1.51316008e+00, -2.60315415e-01,
 -2.62386821e-01],
[ 9.16836886e-01, -3.62176246e-01,  4.78571135e-01,
  1.32509732e-01],
[-7.79513300e-01, -8.22569778e-01,  8.07091462e-02,
  2.64141916e-01],
[-1.02184904e+00, -2.43394714e+00, -1.46640561e-01,
 -2.62386821e-01],
[ 6.86617933e-02, -1.31979479e-01,  2.51221427e-01,
  3.95774101e-01],
[ 1.89829664e-01, -1.97355361e+00,  1.37546573e-01,
 -2.62386821e-01],
[ 3.10997534e-01, -3.62176246e-01,  5.35408562e-01,
  2.64141916e-01],
[-2.94841818e-01, -3.62176246e-01, -8.98031345e-02,
  1.32509732e-01],
[ 1.03800476e+00,  9.82172869e-02,  3.64896281e-01,
  2.64141916e-01],
[-2.94841818e-01, -1.31979479e-01,  4.21733708e-01,
  3.95774101e-01],
[-5.25060772e-02, -8.22569778e-01,  1.94384000e-01,
 -2.62386821e-01],
[ 4.32165405e-01, -1.97355361e+00,  4.21733708e-01,
  3.95774101e-01],
[-2.94841818e-01, -1.28296331e+00,  8.07091462e-02,
 -1.30754636e-01],
[ 6.86617933e-02,  3.28414053e-01,  5.92245988e-01,
  7.90670654e-01],
[ 3.10997534e-01, -5.92373012e-01,  1.37546573e-01,
  1.32509732e-01],
[ 5.53333275e-01, -1.28296331e+00,  6.49083415e-01,
  3.95774101e-01],
[ 3.10997534e-01, -5.92373012e-01,  5.35408562e-01,
  8.77547895e-04],
[ 6.74501145e-01, -3.62176246e-01,  3.08058854e-01,
  1.32509732e-01],
[ 9.16836886e-01, -1.31979479e-01,  3.64896281e-01,
  2.64141916e-01],
[ 1.15917263e+00, -5.92373012e-01,  5.92245988e-01,
  2.64141916e-01],
[ 1.03800476e+00, -1.31979479e-01,  7.05920842e-01,
  6.59038469e-01],
[ 1.89829664e-01, -3.62176246e-01,  4.21733708e-01,
  3.95774101e-01],
[-1.73673948e-01, -1.05276654e+00, -1.46640561e-01,
 -2.62386821e-01],
[-4.16009689e-01, -1.51316008e+00,  2.38717193e-02,
 -1.30754636e-01],
[-4.16009689e-01, -1.51316008e+00, -3.29657076e-02,
 -2.62386821e-01],
[-5.25060772e-02, -8.22569778e-01,  8.07091462e-02,
  8.77547895e-04],
[ 1.89829664e-01, -8.22569778e-01,  7.62758269e-01,
  5.27406285e-01],
[-5.37177559e-01, -1.31979479e-01,  4.21733708e-01,
  3.95774101e-01],
[ 1.89829664e-01,  7.88807586e-01,  4.21733708e-01,
  5.27406285e-01],
[ 1.03800476e+00,  9.82172869e-02,  5.35408562e-01,
  3.95774101e-01],
[ 5.53333275e-01, -1.74335684e+00,  3.64896281e-01,
  1.32509732e-01],
[-2.94841818e-01, -1.31979479e-01,  1.94384000e-01,
  1.32509732e-01],
[-4.16009689e-01, -1.28296331e+00,  1.37546573e-01,
  1.32509732e-01],
[-4.16009689e-01, -1.05276654e+00,  3.64896281e-01,
  8.77547895e-04],
[ 3.10997534e-01, -1.31979479e-01,  4.78571135e-01,
  2.64141916e-01],
[-5.25060772e-02, -1.05276654e+00,  1.37546573e-01,
  8.77547895e-04],
[-1.02184904e+00, -1.74335684e+00, -2.60315415e-01,
 -2.62386821e-01],
[-2.94841818e-01, -8.22569778e-01,  2.51221427e-01,
  1.32509732e-01],
[-1.73673948e-01, -1.31979479e-01,  2.51221427e-01,
  8.77547895e-04],
[-1.73673948e-01, -3.62176246e-01,  2.51221427e-01,
  1.32509732e-01],
[ 4.32165405e-01, -3.62176246e-01,  3.08058854e-01,
  1.32509732e-01],
[-9.00681170e-01, -1.28296331e+00, -4.30827696e-01,
 -1.30754636e-01],
[-1.73673948e-01, -5.92373012e-01,  1.94384000e-01,
  1.32509732e-01],
[ 5.53333275e-01,  5.58610819e-01,  1.27429511e+00,
  1.71209594e+00],
[-5.25060772e-02, -8.22569778e-01,  7.62758269e-01,
  9.22302838e-01],
[ 1.52267624e+00, -1.31979479e-01,  1.21745768e+00,
  1.18556721e+00],
[ 5.53333275e-01, -3.62176246e-01,  1.04694540e+00,
  7.90670654e-01],
[ 7.95669016e-01, -1.31979479e-01,  1.16062026e+00,
  1.31719939e+00],
[ 2.12851559e+00, -1.31979479e-01,  1.61531967e+00,
  1.18556721e+00],
[-1.14301691e+00, -1.28296331e+00,  4.21733708e-01,
  6.59038469e-01],
[ 1.76501198e+00, -3.62176246e-01,  1.44480739e+00,
  7.90670654e-01],
[ 1.03800476e+00, -1.28296331e+00,  1.16062026e+00,
  7.90670654e-01],
[ 1.64384411e+00,  1.24920112e+00,  1.33113254e+00,
  1.71209594e+00],
[ 7.95669016e-01,  3.28414053e-01,  7.62758269e-01,
  1.05393502e+00],
[ 6.74501145e-01, -8.22569778e-01,  8.76433123e-01,
  9.22302838e-01],
[ 1.15917263e+00, -1.31979479e-01,  9.90107977e-01,
  1.18556721e+00],
[-1.73673948e-01, -1.28296331e+00,  7.05920842e-01,
  1.05393502e+00],
[-5.25060772e-02, -5.92373012e-01,  7.62758269e-01,
  1.58046376e+00],
[ 6.74501145e-01,  3.28414053e-01,  8.76433123e-01,
  1.44883158e+00],
[ 7.95669016e-01, -1.31979479e-01,  9.90107977e-01,
  7.90670654e-01],
[ 2.24968346e+00,  1.70959465e+00,  1.67215710e+00,
  1.31719939e+00],
[ 2.24968346e+00, -1.05276654e+00,  1.78583195e+00,
  1.44883158e+00],
[ 1.89829664e-01, -1.97355361e+00,  7.05920842e-01,
  3.95774101e-01],
[ 1.28034050e+00,  3.28414053e-01,  1.10378283e+00,
  1.44883158e+00],
[-2.94841818e-01, -5.92373012e-01,  6.49083415e-01,
  1.05393502e+00],
[ 2.24968346e+00, -5.92373012e-01,  1.67215710e+00,
  1.05393502e+00],
[ 5.53333275e-01, -8.22569778e-01,  6.49083415e-01,
  7.90670654e-01],
[ 1.03800476e+00,  5.58610819e-01,  1.10378283e+00,
  1.18556721e+00],
[ 1.64384411e+00,  3.28414053e-01,  1.27429511e+00,
  7.90670654e-01],
[ 4.32165405e-01, -5.92373012e-01,  5.92245988e-01,
  7.90670654e-01],
[ 3.10997534e-01, -1.31979479e-01,  6.49083415e-01,
  7.90670654e-01],
[ 6.74501145e-01, -5.92373012e-01,  1.04694540e+00,
  1.18556721e+00],
[ 1.64384411e+00, -1.31979479e-01,  1.16062026e+00,
  5.27406285e-01],
[ 1.88617985e+00, -5.92373012e-01,  1.33113254e+00,
  9.22302838e-01],
[ 2.49201920e+00,  1.70959465e+00,  1.50164482e+00,
  1.05393502e+00],
[ 6.74501145e-01, -5.92373012e-01,  1.04694540e+00,
  1.31719939e+00],
[ 5.53333275e-01, -5.92373012e-01,  7.62758269e-01,
  3.95774101e-01],
[ 3.10997534e-01, -1.05276654e+00,  1.04694540e+00,
  2.64141916e-01],
       [ 2.24968346e+00, -1.31979479e-01,  1.33113254e+00,
         1.44883158e+00],
       [ 5.53333275e-01,  7.88807586e-01,  1.04694540e+00,
         1.58046376e+00],
       [ 6.74501145e-01,  9.82172869e-02,  9.90107977e-01,
         7.90670654e-01],
       [ 1.89829664e-01, -1.31979479e-01,  5.92245988e-01,
         7.90670654e-01],
       [ 1.28034050e+00,  9.82172869e-02,  9.33270550e-01,
         1.18556721e+00],
       [ 1.03800476e+00,  9.82172869e-02,  1.04694540e+00,
         1.58046376e+00],
       [ 1.28034050e+00,  9.82172869e-02,  7.62758269e-01,
         1.44883158e+00],
       [-5.25060772e-02, -8.22569778e-01,  7.62758269e-01,
         9.22302838e-01],
       [ 1.15917263e+00,  3.28414053e-01,  1.21745768e+00,
         1.44883158e+00],
       [ 1.03800476e+00,  5.58610819e-01,  1.10378283e+00,
         1.71209594e+00],
       [ 1.03800476e+00, -1.31979479e-01,  8.19595696e-01,
         1.44883158e+00],
       [ 5.53333275e-01, -1.28296331e+00,  7.05920842e-01,
         9.22302838e-01],
       [ 7.95669016e-01, -1.31979479e-01,  8.19595696e-01,
         1.05393502e+00],
       [ 4.32165405e-01,  7.88807586e-01,  9.33270550e-01,
         1.44883158e+00],
       [ 6.86617933e-02, -1.31979479e-01,  7.62758269e-01,
         7.90670654e-01]])
```

Hemos normalizado con **One hot**, pero esta matriz esta acomodada de una manera no adecuada, así que tenemos que aplicar una transposición de los datos.

In [ ]:

```
data_scaled.T

#definiendo matriz de co-varianza
covarianza_matriz=np.cov(data_scaled.T)

#Esto me deberia de dar una matriz 4 x 4
covarianza_matriz
```

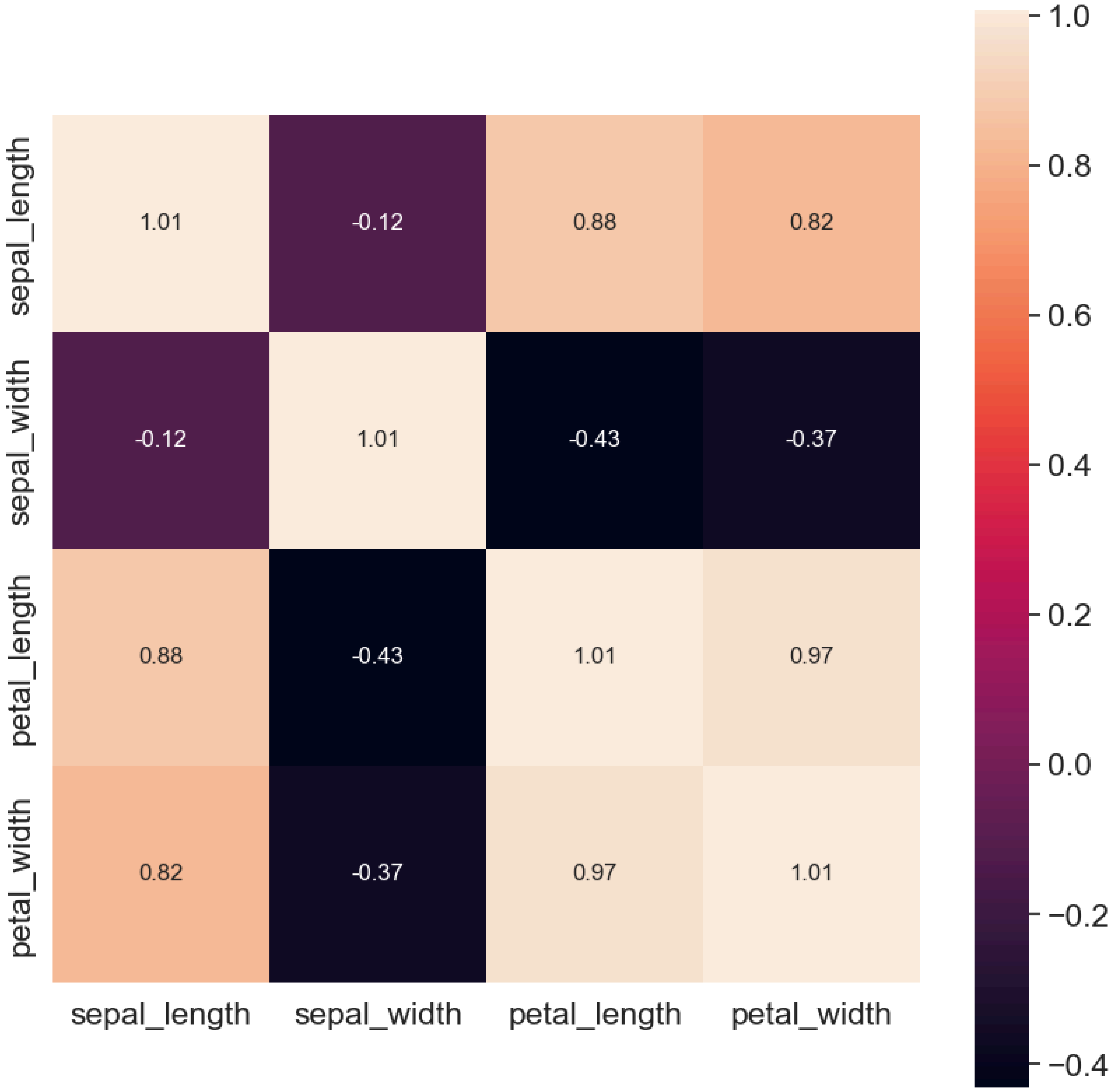
Out[ ]:

```
array([[ 1.00671141, -0.11835884,  0.87760447,  0.82343066],
       [-0.11835884,  1.00671141, -0.43131554, -0.36858315],
       [ 0.87760447, -0.43131554,  1.00671141,  0.96932762],
       [ 0.82343066, -0.36858315,  0.96932762,  1.00671141]])
```

La diagonal principal es la que tiene la relación perfecta de co-relación, hay otra forma de verla mediante la gráfica de **heatmap**

In [ ]:

```
plt.figure(figsize=(10,10))
sns.set(font_scale=1.5)
hm = sns.heatmap(covarianza_matriz,
                 cbar=True,
                 annot=True,
                 square=True,
                 fmt='.2f',
                 annot_kws={'size': 12},
                 yticklabels=['sepal_length', 'sepal_width', 'petal_length', 'petal_width'],
                 xticklabels=['sepal_length', 'sepal_width', 'petal_length', 'petal_width'])
```



Podemos ver como esta mejor afianzado de manera gráfica, para poder ver la relación entre las variables.

Resumen

La idea de esto es que mediante esta matriz veamos la relación de las variables para poder seleccionar unas para el modelo de MAchine Learning y no necesito pasar todas, ese es el proceso que llamamos **Reducción de datos**, que esta basado en la matriz de covarianza y se le llama **ANALISIS DE COMPONENTES PRINCIPALES**: Es una técnica que involucra Algebra Lineal, manejo de vectores.
En las siguientes clases vamos a ver de que trata

Extra

HeatMap